A KNN-Based Non-Parametric Conditional Independence Test for Mixed Data and Application in Causal Discovery

Johannes Huegle(✉), Christopher Hagedorn, and Rainer Schlosser

University of Potsdam, Hasso Plattner Institute, Potsdam, Germany
{Johannes.Huegle,Christopher.Hagedorn,Rainer.Schlosser}@hpi.de

Abstract. Testing for Conditional Independence (CI) is a fundamental task for causal discovery but is particularly challenging in mixed discrete-continuous data. In this context, inadequate assumptions or discretization of continuous variables reduce the CI test's statistical power, which yields incorrect learned causal structures. In this work, we present a non-parametric CI test leveraging k-nearest neighbor (kNN) methods that are adaptive to mixed discrete-continuous data. In particular, a kNN-based conditional mutual information estimator serves as the test statistic, and the p-value is calculated using a kNN-based local permutation scheme. We prove the CI test's statistical validity and power in mixed discrete-continuous data, which yields consistency when used in constraint-based causal discovery. An extensive evaluation of synthetic and real-world data shows that the proposed CI test outperforms state-of-the-art approaches in the accuracy of CI testing and causal discovery, particularly in settings with low sample sizes.

Keywords: Non-Parametric CI Testing · Causal Discovery · Mixed Data

1 Introduction

Conditional Independence (CI) testing is at the core of causal discovery (Sect. 1.1), but particularly challenging in many real-world scenarios (Sect. 1.2). Therefore, we propose a data-adaptive CI test for mixed discrete-continuous data (Sect. 1.3).

1.1 Conditional Independence in Causal Discovery

Causal discovery has received widespread attention as the knowledge of underlying causal structures improves decision support within many real-world scenarios [17,46]. For example, in discrete manufacturing, causal discovery is the key to root cause analysis of failures and quality deviations, cf. [25].

D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14169, pp. 541–558, 2023.
https://doi.org/10.1007/978-3-031-43412-9_32

Causal structures between a finite set of random variables $\mathbf{V} = \{X, Y, \dots\}$ are encoded in a Causal Graphical Model (CGM) consisting of a Directed Acyclic Graph (DAG) $\mathcal{G}$, and the joint distribution over the variables $\mathbf{V}$, denoted by $P_{\mathbf{V}}$, cf. [38,46]. In $\mathcal{G}$, a directed edge $X \to Y$ depicts a direct causal mechanism between the two respective variables X and Y, for $X, Y \in \mathbf{V}$. Causal discovery aims to derive as many underlying causal structures in $\mathcal{G}$ from observational data as possible building upon the coincidence between the causal structures of $\mathcal{G}$ and the CI characteristics of $P_{\mathbf{V}}$ [46]. Therefore, constraint-based methods, such as the well-known PC algorithm, apply CI tests to recover the causal structures, cf. [8]. For instance, if a CI test states the conditional independence of variables X and Y given a (possibly empty) set of variables $Z \subseteq \mathbf{V} \setminus \{X, Y\}$, denoted by $X \perp\!\!\!\perp Y \mid Z$, then there is no edge between X and Y. Constraint-based methods are flexible and exist in various extensions, e.g., to allow for latent variables or cycles [42,46,47], or are used for causal feature selection [50]. Hence, they are popular in practice [33].

1.2 Challenges in Practice

In principle, constraint-based methods do not make any assumption on the functional form of causal mechanisms or parameters of the joint distribution. However, they require access to a CI oracle that captures all CI characteristics such that selecting an appropriate CI test is fundamental and challenging [17,33]. In practice, the true statistical properties are mostly unknown such that inadequate assumptions, e.g., of parametric CI tests, yield incorrect learned causal structures [46]. For example, the well-known partial Pearson's correlation-based CI test via Fisher's Z transformation assumes that $P_{\mathbf{V}}$ is multivariate Gaussian [3,27]. Hence, the underlying causal mechanisms are assumed to be linear and conditional independence cannot be detected if the mechanisms are non-linear. Further, the omnipresence of mixed discrete-continuous data, e.g., continuous quality measurements and discrete failure messages in discrete manufacturing [20], impedes the selection of appropriate CI tests in real-world scenarios [19,33]. In this case, parametric models that allow for mixed discrete-continuous data usually make further restrictions, such as conditional Gaussian models assuming that discrete variables have discrete parents only [40]. Hence, for simplification in practice, continuous variables are often discretized to use standard CI tests such as Pearson's χ^2 test for discrete data, cf. [20,23,35], to the detriment of the accuracy of the learned causal structures [12,40].

1.3 Contribution and Structure

In this work, we propose `mCMIkNN`[1], a data-adaptive CI test for mixed discrete-continuous data and its application to causal discovery. Our contributions are:

- We propose a kNN-based local conditional permutation scheme to derive a non-parametric CI test using a kNN-based CMI estimator as a test statistic.
- We provide theoretical results on the CI test's validity and power. In particular, we prove that `mCMIkNN` is able to control type I and type II errors.

[1] Code and Appendix can be found on https://github.com/hpi-epic/mCMIkNN.

– We show that `mCMIkNN` allows for consistent estimation of causal structures when used in constraint-based causal discovery.
– An extensive evaluation on synthetic and real-world data shows that `mCMIkNN` outperforms state-of-the-art competitors, particularly for low sample sizes.

The remainder of this paper is structured as follows. In Sect. 2, we examine the problem of CI testing and related work. In Sect. 3, we provide background on kNN-based CMI estimation. In Sect. 4, we introduce `mCMIkNN` and prove theoretical results. In Sect. 5, we empirically evaluate the accuracy of our CI test `mCMIkNN` compared to state-of-the-art approaches. In Sect. 6, we conclude our work.

2 Conditional Independence Testing Problem

In this section, we provide a formalization of the CI testing problem (Sect. 2.1) together with existing fundamental limits of CI testing (Sect. 2.2) before considering related work on CI testing for mixed discrete-continuous data (Sect. 2.3).

2.1 Problem Description

Let $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{B}, P_{XYZ})$ be a probability space defined on the metric space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with dimensionality $d_X + d_Y + d_Z$, equipped with the Borel σ-algebra $\mathcal{B}$, and a regular joint probability measure P_{XYZ}. Hence, we assume that the d_X, d_Y, and d_Z-dimensional random variables X, Y, and Z take values in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ according to the marginal mixed discrete-continuous probability distributions P_X, P_Y, and P_Z. I.e., single variables in X, Y, or Z may follow a discrete, a continuous, or a mixture distribution.

We consider the problem of testing the CI of two random vectors X and Y given a (possibly empty) random vector Z sampled according to the mixed discrete-continuous probability distribution P_{XYZ}, i.e., testing the null hypothesis of CI $H_0 : X \perp\!\!\!\perp Y \mid Z$ against the alternative hypothesis of dependence $H_1 : X \not\perp\!\!\!\perp Y \mid Z$. Therefore, let $(x_i, y_i, z_i)_{i=1}^n$ be n i.i.d. observations sampled from P_{XYZ} such that we aim to derive a CI test $\Phi_n : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \times [0,1] \to \{0,1\}$ that rejects H_0 if $\Phi_n = 1$ given a nominal level $\alpha \in [0,1]$.

2.2 Fundamental Limits of CI Testing

The general problem of CI testing is extensively studied, as it is a fundamental concept beyond its application in constraint-based causal discovery [11]. In this context, it is necessary to note that Shah and Peters [45] provided a *no-free lunch theorem* for CI that, given a continuously distributed conditioning set Z, it is impossible to derive a CI test that is able to control the type I error, via for instance a permutation scheme, and has nontrivial power without additional restrictions. But, under the restriction that the conditional distribution $P_{X|Z}$ is known or can be approximated sufficiently, conditional permutation (CP) tests can calibrate a test statistic guaranteeing a controlled type I error [4]. Further, the recent work of Kim et al. [28] shows that the problem of CI testing is more generally determined by the probability of observing *collisions* in Z.

2.3 Related Work

We consider the problem of CI testing and its application in causal discovery. In this context, constraint-based methods require CI tests that (R1) yield accurate CI decisions, and (R2) are computationally feasible as they are applied hundreds of times. Generally, CI testing for mixed discrete-continuous data can be categorized into discretization-based, parametric, and non-parametric approaches.

Discretization-Based Approaches: As CI tests for discrete variables are well-studied, continuous variables are often discretized, cf. [23,35]. In this context, commonly used CI tests for discrete data are Pearson's $\mathcal{X}^2$ and likelihood ratio tests [13,39,46]. Although discretization simplifies the testing problem, the resulting information loss yields a decreased accuracy [12,40], cf. (R1).

Parametric CI Testing: Postulating an underlying parametric functional model allows for a regression-based characterization of CI that can be used to construct valid CI tests. Examples are well-known likelihood ratio tests, e.g., assuming conditional Gaussianity (CG) [1,44] or using multinomial logistic regression models [48]. Another stream of research focuses on Copula models to examine CI characteristics in mixed discrete-continuous data, where variables are assumed to be induced by latent Gaussian variables such that CI can be determined by examining the correlation matrix of the latent variables model [9,10]. As these approaches require that the postulated parametric models hold, they may yield invalid CI decisions if assumptions are inaccurate [46], cf. (R1).

Non-Parametric CI Testing: Non-parametric CI testing faces the twofold challenge to, first, derive a test statistic from observational data without parametric assumptions, and second, derive the p-value given that the test statistic's distribution under H_0 may be unknown. In continuous data, a wide range of methods is used for non-parametric CI testing, as reviewed by Li and Fan [32]. For example, kernel-based approaches, such as `KCIT` [52], test for vanishing correlations within Reproducing Kernel Hilbert Spaces (RKHS). Another example is `CMIknn` from Runge [43], which uses a kNN-based estimator to test for a vanishing Conditional Mutual Information (CMI) in combination with a local permutation scheme. The recent emergence of non-parametric CMI estimators for mixed discrete-continuous data provides the basis for new approaches to non-parametric CI testing. For example, the construction of adaptive histograms derived following the minimum description length (MDL) principle allows for estimating CMI from mixed discrete-continuous data [6,34,36,51]. In this case, CMI can be estimated via discrete plug-in estimators as the data is adaptively discretized according to the histogram with minimal MDL. Hence, the estimated test statistic follows the common $\mathcal{X}^2$ distribution, which allows for derivation via Pearson's $\mathcal{X}^2$ test, aHisχ^2, see [36]. However, MDL approaches suffer from their worst-case computational complexity and weaknesses regarding a low number of samples, cf. (R2). Another approach for non-parametric CMI estimation builds

upon kNN methods, which are well-studied in continuous data, cf. [15,29,30], and have recently been applied to mixed discrete-continuous data [16,37]. As the asymptotic distribution of kNN-based estimators is unclear, it remains to show that they can be used as a test statistic for a valid CI. In this context, it is worth noticing that permutation tests yield more robust constraint-based causal discovery than asymptotic CI tests, particularly for small sample sizes [49], cf. (R1). Following this, we combine a kNN-based CMI estimator and a kNN based local CP scheme (similar to Runge [43], which is restricted to the continuous case), and additionally provide theoretical results on the test's validity and power.

3 Background: KNN-Based CMI Estimation

In this section, we provide information on kNN-based CMI estimation for mixed discrete-continuous data (Sect. 3.1). Further, we introduce an algorithmic description of the estimator (Sect. 3.2) and recap theoretical results (Sect. 3.3).

3.1 Introduction to CMI Estimation

A commonly used test statistic is the Conditional Mutual Information (CMI) $I(X;Y|Z)$ as it provides a general measure of variables' CI, i.e., $I(X;Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y \,|\, Z$, see [16,18,43]. Generally, $I(X;Y|Z)$ is defined as $I(X;Y|Z) = \int \log\left(\frac{dP_{XY|Z}}{d(P_{X|Z}\times P_{Y|Z})}\right) dP_{XYZ}$, where $\frac{dP_{XY|Z}}{d(P_{X|Z}\times P_{Y|Z})}$ is the Radon-Nikodym derivative of the joint conditional measure, $P_{XY|Z}$, with respect to the product of the marginal conditional measures, $P_{X|Z} \times P_{Y|Z}$. Note the non-singularity of P_{XYZ} ensures the existence of a product reference measure and that the Radon-Nikodym derivative is well-defined [37, Lem. 2.1, Thm. 2.2]. Although well-defined, estimating CMI $I(X;Y|Z)$ from mixed discrete-continuous data is a particularly hard challenge [16,36,37]. Generally, CMI estimation can be tackled by expressing $I(X;Y|Z)$ in terms of Shannon entropies, i.e., $I(X;Y|Z) = H(X,Y,Z) - H(X,Z) - H(Y,Z) + H(Z)$ with Shannon entropy $H(W)$ for all cases $W = XYZ, XZ, YZ, Z$, respectively, cf. [18,36,37]. In the continuous case, the KSG technique from Kraskov et al. [30] estimates the Shannon entropy $H(W)$ locally for every sample $(w_i)_{i=1}^n$ where $w_i \sim P_W$, i.e., estimating $H(W)$ via $\widehat{H}_n(W) = -\sum_{i=1}^n \log \widehat{f_W(w_i)}$ by considering the k-nearest neighbors within the ℓ_∞-norm for every sample $i = 1, ..., n$ to locally estimate the density f_W density of $W = XYZ, XZ, YZ, Z$, respectively, cf. [18,36,37]. For mixed discrete-continuous data, there is a non-zero probability that the kNN distance is zero for some samples. In this case, Gao et al. [16] extended the KSG technique by fixing the radius and using a plug-in estimator that differentiates between mixed, continuous, and discrete points. Recently, Mesner and Shalizi [37] extended this idea to derive a consistent CMI estimator in the mixed discrete-continuous case.

Algorithm 1. kNN-based CMI Estimator [37]

Input: Samples $(x, y, z) := (x_i, y_i, z_i)_{i=1}^n$, and kNN-parameter k_{CMI}
Output: The estimated value $\hat{I}_n(x; y|z)$ of the CMI $I(X; Y|Z)$

1: Let $d_{i,j}(w) := \|(w_i) - (w_j)\|_\infty$ for $w \subseteq (x, y, z)$, $i, j = 1, \ldots, n$
2: **for** $i = 1, \ldots, n$ **do**
3: $\quad \rho_i :=$ the k_{CMI}-smallest distance in $\{d_{i,j}(x, y, z), j \neq i\}$ ▷ Adapt k_{CMI} acc. ρ_i
4: $\quad \tilde{k}_i := |\{(x_j, y_j, z_j) : d_{i,j}(x, y, z) \leq \rho_i, j \neq i\}|$
5: $\quad n_{xz,i} := |\{(x_j, z_j) : d_{i,j}(x, z) \leq \rho_i, j \neq i\}|$ ▷ Local estimates
6: $\quad n_{yz,i} := |\{(y_j, z_j) : d_{i,j}(y, z) \leq \rho_i, j \neq i\}|$
7: $\quad n_{z,i} := |\{(z_j : d_{i,j}(z) \leq \rho_i, j \neq i\}|$
8: $\quad \xi_i := \psi(\tilde{k}_i) - \psi(n_{xz,i}) - \psi(n_{yz,i}) + \psi(n_{z,i})$
9: **end for**
10: $\hat{I}_n(x; y|z) = \frac{1}{n} \sum_{i=1}^n \xi_i$ ▷ Global CMI estimation
11: **return** $\max(\hat{I}_n(x; y|z), 0)$

3.2 Algorithm for KNN-Based CMI Estimation

Algorithm 1 provides an algorithmic description of the theoretically examined estimator $\hat{I}_n(X; Y|Z)$ developed by Mesner and Shalizi [37]. The basic idea is to take the mean of Shannon entropies estimated locally for each sample $i = 1, ..., n$ considering samples $j \neq i$, $j = 1, ..., n$, that are close to i according to the ℓ_∞-norm, i.e., under consideration of the respective sample distance $d_{i,j}(w) := \|(w_i) - (w_j)\|_\infty$, $i, j = 1, ..., n$, of $w = (w_i)_{i=1}^n$ for all cases $w = xyz, xy, yz, z$ (see Algorithm 1, line 1). In this context, fixation of a kNN radius ρ_i used for local estimation of Shannon entropies yields a consistent global estimator. Therefore, for each sample $i = 1, \ldots, n$, let ρ_i be the smallest distance between (x_i, y_i, z_i) and the k_{CMI}-nearest sample (x_j, y_j, z_j), $j \neq i, j = 1, \ldots, n$, and replace k_{CMI} with $\tilde{k}_i$, the number of samples whose distance to (x_i, y_i, z_i) is smaller or equal to ρ_i (see Algorithm 1, line 3-4). For discrete or mixed discrete-continuous samples $(x_i, y_i, z_i)_{i=1}^n$ it holds that $\rho_i = 0$, and there may be more samples than k_{CMI} samples with zero distance. In this case, adapting the number of considered samples $\tilde{k}_i$ to all samples with zero distance prevents undercounting, which, otherwise, yields a bias of the CMI estimator, see [37]. In case of continuous samples $(x_i, y_i, z_i)_{i=1}^n$, there are exactly $\tilde{k}_i = k_{CMI}$ samples within the k_{CMI}-nearest distance with probability 1. The next step estimates the Shannon entropies required by the $3H$-principle locally for each sample i, $i = 1, \ldots, n$. Therefore, let $n_{xz,i}, n_{yz,i}$, and $n_{z,i}$ be the numbers of $\tilde{k}_i$-nearest samples within the distance of ρ_i in the respective subspace XZ, YZ, and Z (see Algorithm 1, lines 5-7). Fixing the local kNN distance ρ_i, using the ℓ_∞-norm, simplifies the local estimation as most relevant terms for CMI estimation using the $3H$-principle cancel out, i.e., $\xi_i := -\widehat{f_{XYZ}(x_i, y_i, z_i)} + \widehat{f_{XZ}(x_i, z_i)} + \widehat{f_{YZ}(y_i, z_i)} - \widehat{f_Z(z_i)} = \psi(\tilde{k}_i) - \psi(n_{xz,i}) - \psi(n_{yz,i}) + \psi(n_{z,i})$, with digamma function ψ (see Algorithm 1, line 8) [16,37]. Then, the global CMI estimate $\hat{I}_n(x; y|z)$ is the average of the local CMI estimates ξ_i of each sample $(x_i, y_i, z_i)_{i=1}^n$, and the positive part is returned, as CMI or MI are non-negative (see Algorithm 1, line 10-11).

3.3 Properties of KNN-Based CMI Estimation

We recap the theoretic results of $\hat{I}_n(X, Y|Z)$ proved by Mesner and Shalizi [37]. Under mild assumptions, $\hat{I}_n(x; y|z)$ is asymptotically unbiased, see [37, Thm. 3.1].

Corollary 1 ***(Asymptotic-Unbiasedness of*** $\hat{I}_n(x; y|z)$ *[37, Thm. 3.1])* *Let* $(x_i, y_i, z_i)_{i=1}^n$ *be i.i.d. samples from* P_{XYZ}*. Assume*

(A1) $P_{XY|Z}$ *is non-singular such that* $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ *is well-defined, and assume, for some* $C > 0$, $f(x, y, z) < C$ *for all* $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$*;*
(A2) $\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : P_{XYZ}((x, y, z)) > 0\}$ *countable and nowhere dense in* $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$*;*
(A3) $k_{CMI} = k_{CMI,n} \to \infty$ *and* $\frac{k_{CMI,n}}{n} \to 0$ *as* $n \to \infty$*;*

then $\mathbb{E}_{P_{XYZ}}\left[\hat{I}_n(x; y|z)\right] \to I(X; Y|Z)$ *as* $n \to \infty$.

While (A1) seems rather technical, checking for non-singularity is helpful for data analysis by checking sufficient conditions. Given non-singularity, assumptions (A2) and (A3) are satisfied whenever P_{XYZ} is (i) (finitely) discrete, (ii) continuous, (iii) some dimensions are (countably) discrete and some are continuous, and (iv) a mixture of the previous cases, which covers most real-world data. For more details on the assumptions, see Appendix A.

We prove that the CMI estimator $\hat{I}_n(X; Y|Z)$ described in Algorithm 1 is consistent.

Corollary 2 ***(Consistency of*** $\hat{I}_n(x; y|z)$***)*** *Let* $(x_i, y_i, z_i)_{i=1}^n$ *be i.i.d. samples from* P_{XYZ} *and assume (A1)-(A3) of Cor. 1 hold. Then, for all* $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}_{P_{XYZ}}\left(\left|\hat{I}_n(x; y|z) - I(X; Y|Z)\right| > \epsilon\right) = 0$.

Proof. Recap that $\hat{I}_n(x; y|z)$ has asymptotic vanishing variance [37, Thm. 3.2], i.e., $\lim\limits_{n\to\infty} \mathrm{Var}(\hat{I}_n(x; y|z)) = 0$, and is asymptotically unbiased, see Cor. 1 or [37, Thm. 3.1]. The consistency of $\hat{I}_n(x; y|z)$ follows from Chebyshev's inequality. □

Therefore, the kNN-based estimator described in Algorithm 1 serves as a valid test statistic for $H_0 : X \perp\!\!\!\perp Y \mid Z$ vs. $H_1 : X \not\perp\!\!\!\perp Y \mid Z$. Note that, $\hat{I}_n(x; y|z)$ is biased towards zero for high-dimensional data with fixed sample size, i.e., it suffers from the curse of dimensionality, see [37, Thm. 3.3].

Corollary 3 ***(Dimensionality-Biasedness of*** $\hat{I}_n(x; y|z)$ *[37, Thm. 3.3])* *Let* $(x_i, y_i, z_i)_{i=1}^n$ *be i.i.d. samples from* P_{XYZ} *and assume (A1)-(A3) of Cor. 1 hold, if the entropy rate of* Z *is nonzero, i.e.,* $\lim\limits_{d_Z\to\infty} \frac{1}{d_Z} H(Z) \neq 0$*, then, for fixed dimensions* d_X *and* d_Y, $\mathbb{P}_{P_{XYZ}}\left(\hat{I}_n(x; y|z) = 0\right) \to 1$ *as* $d_Z \to \infty$.

Hence, even with asymptotic consistency, one must pay attention when estimating $\hat{I}_n(X; Y|Z)$ in high-dimensional settings, particularly for low sample sizes.

4 mCMIkNN: Our Approach on Non-Parametric CI Testing

In this section, we recap the concept of Conditional Permutation (CP) schemes for CI testing (Sect. 4.1). Then, we introduce our approach for kNN-based CI testing in mixed discrete-continuous data, called mCMIkNN (Sect. 4.2). We prove that mCMIkNN is able to control type I and type II errors (Sect. 4.3). Moreover, we examine mCMIkNN-based causal discovery and prove its consistency (Sect. 4.4).

4.1 Introduction to Conditional Permutation Schemes

Using permutation schemes for non-parametric independence testing between two variables X and Y has a long history in statistics, cf. [5,22,31]. The basic idea is to compare an appropriate test statistic for independence calculated from the original samples $(x_i, y_i)_{i=1}^n$ against the test statistics calculated M_{perm} times from samples $(x_{\pi_m(i)}, y_i)_{i=1}^n$ for a permutation π_m of $\{1, \ldots, n\}$, $m = 1, \ldots, M_{perm}$, i.e., where samples of X are randomly permuted such that $H_0 : X \perp\!\!\!\perp Y$ holds. In the discrete case, a permutation scheme to test for CI, i.e., for $H_0 : X \perp\!\!\!\perp Y \,|\, Z$, can be achieved by permuting X for each realization $Z = z$ to utilize the unconditional $X \perp\!\!\!\perp Y \,|\, Z = z$. In contrast, testing for CI in continuous or mixed discrete-continuous data is more challenging [45], as simply permuting X without considering the confounding effect of Z may yield very different marginal distributions, hence, suffers in type I error control [4,28]. Therefore, Conditional Permutation (CP) schemes aim to compare a test statistic estimated from the original data $(x_i, y_i, z_i)_{i=1}^n$, with test statistics estimated from, conditionally on Z, permuted samples $(x_{\pi_m(i)}, y_i, z_i)_{i=1}^n$, $m = 1, ..., M_{perm}$ to ensure $H_0 : X \perp\!\!\!\perp Y \,|\, Z$. Then, the $M_{perm} + 1$ samples $(x_i, y_i, z_i)_{i=1}^n$ and $(x_{\pi_m(i)}, y_i, z_i)_{i=1}^n$, $m = 1, ..., M_{perm}$ are exchangeable under H_0, i.e., are drawn with replacement such that the p-value can be calculated in line with common Monte Carlo simulations [4,28]. This requires either an approximation of $P_{X|Z}$ either based upon model assumptions to simulate $P_{X|Z}$ [4], or using an adaptive binning strategy of Z such that permutations can be drawn for each binned realization $Z = z$ [28] (both focusing on the continuous case). To provide a data-adaptive approach valid in mixed discrete-continuous data without too restrictive assumptions, cf. (R1), which is computationally feasible, cf. (R2), we propose a local CP scheme leveraging ideas of kNN-based methods, cf. Section 3. In particular, our local CP scheme draws samples $(x_{\pi_m(i)}, y_i, z_i)_{i=1}^n$ such that (I) the marginal distributions are preserved, and (II) x_i is replaced by $x_{\pi_m(i)}$ only locally regarding the k_{perm}-nearest distance σ_i in the space of Z. Intuitively, the idea is similar to common conditional permutation schemes in the discrete case, where entries of the variable X are permuted for each realization $Z = z$, but considering local permutations regarding the neighborhood of $Z = z$.

4.2 Algorithm for KNN-Based CI Testing

Algorithm 2 gives an algorithmic description of our kNN-based local CP scheme for non-parametric CI testing in mixed discrete-continuous data.

Algorithm 2. `mCMIkNN`: kNN-based non-parametric CI Test

Input: Samples $(x, y, z) := (x_i, y_i, z_i)_{i=1}^n$, and parameters k_{CMI}, k_{perm}, and M_{perm}
Output: The estimated p-value $p_{perm,n}$ for $H_0 : X \perp\!\!\!\perp Y \mid Z$

1: $\hat{I}_n := \hat{I}_n(x; y|z)$
2: **for** $i = 1, \ldots, n$ **do** ▷ Neighbors within k_{perm}NN-distance σ_i in Z
3: $\sigma_i :=$ k_{perm} smallest distance in $\{\|(z_i) - (z_j)\|_\infty, j \neq i$, for $i, j = 1, ..., n\}$
4: $\tilde{\mathbf{z}}_i := \{j : \|(z_i) - (z_j)\|_\infty \leq \sigma_i, j \neq i\}$
5: **end for**
6: **for** $m = 1, \ldots, M_{perm}$ **do** ▷ Local CP scheme
7: $\pi_m^i :=$ permutation of $\tilde{\mathbf{z}}_i, i = 1, \ldots, n$
8: $\pi_m := \pi_m^1 \circ \cdots \circ \pi_m^n$;
9: $\hat{I}_n^{(m)} := \hat{I}_n\left(x^{(m)}; y|z\right)$ where $x^{(m)} := (x_{\pi_m(i)})_{i=1}^n$
10: **end for**
11: $p_{perm,n} := \frac{1}{1+M_{perm}}\left(1 + \sum_{m=1}^{M_{perm}} \mathbb{1}\{\hat{I}_n^{(m)} \geq \hat{I}_n\}\right)$ ▷ Monte Carlo p-value
12: **return** $p_{perm,n}$

First, the sample CMI $\hat{I}_n := \hat{I}_n(x; y|z)$ is estimated from the original samples via Algorithm 1 with parameter k_{CMI} (see Algorithm 2, line 1). To receive local conditional permutations for each sample $(x_i, y_i, z_i)_{i=1}^n$, the k_{perm}-nearest neighbor distance σ_i w.r.t. the ℓ_∞-norm of the subspace of Z is considered. Hence, $\tilde{\mathbf{z}}_i$ is the respective set of indices $j \neq i$, $j = 1, ..., n$ of points with distance smaller or equal to σ_i in the subspace of Z (see Algorithm 2, lines 3-4). According to a Monte Carlo procedure, samples are permuted M_{perm} times (see Algorithm 2, line 6). For each $m = 1, \ldots, M_{perm}$, the local conditional permutation π_m^i, $i = 1, \ldots, n$, is a random permutation of the index set of $\tilde{\mathbf{z}}_i$ such that the global permutation scheme π_m of the samples' index set $\{1, \ldots, n\}$ is achieved by concatenating all local permutations, i.e., $\pi_m := \pi_m^1 \circ ... \circ \pi_m^n$ (see Algorithm 2, lines 7-8). In the case of discrete data, $\tilde{\mathbf{z}}_i$ contains all indices of samples j with distance $\rho_i = 0$ to z_i, i.e., the permutation scheme coincides with discrete permutation tests where permutations are considered according to $Z = z_i$. In the continuous case, $\tilde{\mathbf{z}}_i$ contains exactly the, in space Z, k_{perm}-nearest neighbors' indices and the global permutation scheme approximates $P_{X|Z=z_i}$ locally within k_{perm}-NN distance σ_i of z_i. Therefore, local conditional permuted samples $(x_{\pi_m(i)}, y_i, z_i)$ are drawn by shuffling the values of x_i according to π_m and respective CMI values $\hat{I}_n^{(m)} := \hat{I}_n\left(x^{(m)}; y|z\right)$ are estimated using Algorithm1 (see Algorithm 2, line 9). Hence, by construction, $(x_{\pi_m(i)}, y_i, z_i)$ are drawn under $H_0 : X \perp\!\!\!\perp Y \mid Z$ such that the p-value $p_{perm,n}$ can be calculated according to a Monte Carlo scheme comparing the samples' CMI value $\hat{I}_n$ with the H_0 CMI values $\hat{I}_n^{(m)}$ (see Algorithm 2, line 11).

We define the CI test `mCMIkNN` as $\Phi_{perm,n} := \mathbb{1}\{p_{perm,n} \leq \alpha\}$ for the $p_{perm,n}$ returned by Algorithm 2 and, hence, reject $H_0 : X \perp\!\!\!\perp Y \mid Z$ if $\Phi_n = 1$. The computational complexity of `mCMIkNN` is determined by the kNN searches in Algorithms 1 and 2, which is implemented in $\mathcal{O}(n \times log(n))$ using k-d-trees. For more details on assumptions, parameters, and computational complexity, see Appendix A.

4.3 Properties of mCMIkNN

The following two theorems show that mCMIkNN is valid, i.e., is able to control type I errors, and has non-trivial power, i.e., is able to control type II errors.

Theorem 1 ***(Validity: Type I Error Control of*** $\Phi_{perm,n}$***)***
Let $(x_i, y_i, z_i)_{i=1}^n$ *be i.i.d. samples from* P_{XYZ}*, and assume (A1), (A2), and*

(A4) $k_{perm} = k_{perm,n} \to \infty$ *and* $\frac{k_{perm,n}}{n} \to 0$ *as* $n \to \infty$*,*

hold, then $\Phi_{perm,n}$ *with p-value estimated according to Algorithm 2 is able to control type I error, i.e., for any desired nominal value* $\alpha \in [0, 1]$*, when* H_0 *is true, then*

$$\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[\Phi_{perm,n}] \leq \alpha. \tag{1}$$

Note that this holds true independent of the test statistic $T_n : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \mathbb{R}$. The idea of the proof is to bound the type I error using the total variation distance between the samples' conditional distribution $P^n_{X|Z}$ and the conditional distribution $\widetilde{P}^n_{X|Z}$, approximated by the local CP scheme to simulate H_0 and show that it vanishes for $n \to \infty$. For a detailed proof, see Appendix B.

Theorem 2 ***(Power: Type II Error Control of*** $\Phi_{perm,n}$***)***
Let $(x_i, y_i, z_i)_{i=1}^n$ *be i.i.d. samples from* P_{XYZ}*, and assume (A1) - (A4) hold. Then* $\Phi_{perm,n}$*, with p-value estimated according to Algorithm 2, is able to control type II error, i.e., for any desired nominal value* $\beta \in \left[\frac{1}{1+M_{perm}}, 1\right]$*, when* H_1 *is true, then*

$$\lim_{n\to\infty} \mathbb{E}_{P_{XYZ}}[1 - \Phi_{perm,n}] = 0. \tag{2}$$

Hence, mCMIkNN's power is naturally bounded according to M_{perm}, i.e., $1 - \beta \leq 1 - \frac{1}{1+M_{perm}}$. The proof follows from the asymptotic consistency of $\hat{I}_n(x; y|z)$ and that the local CP scheme allows asymptotic consistent approximating $P_{X|Z}$. For a detailed proof, see Appendix B. Therefore, our work is in line with the result of Shah and Peters [45] and Kim et al. [28] by demonstrating that, under the mild assumptions (A1) and (A2) which allow approximating $P_{X|Z}$, one can derive a CI test that is valid (see Thm. 1), and has non-trivial power (see Thm. 2).

4.4 mCMIkNN-based Constraint-based Causal Discovery

We examine the asymptotic consistency of mCMIkNN-based causal discovery, in particular, using the well-known PC algorithm [46]. Note that constraint-based methods for causal discovery cannot distinguish between different DAGs $\mathcal{G}$ in the same equivalence class. Hence, the PC algorithm aims to find the Completed Partially Directed Acyclic Graph (CPDAG), denoted with $\mathcal{G}_{CPDAG}$, that represents the Markov equivalence class of the true DAG $\mathcal{G}$. Constraint-based methods apply CI tests to test whether $X \perp\!\!\!\perp Y \,|\, Z$ for $X, Y \in \mathbf{V}$ with $d_X = d_Y = 1$, and

$Z \in \mathbf{V} \setminus \{X, Y\}$ iteratively with increasing d_Z given a nominal value α to estimate the undirected skeleton of $\mathcal{G}$ and corresponding separation sets in the first step. In a second step, orienting as many of the undirected edges through the repeated application of deterministic orientation rules yields $\hat{\mathcal{G}}_{CPDAG}(\alpha)$ [26,46].

Theorem 3 ***(Consistency of mCMIkNN-based Causal Discovery)*** *Let* $\mathbf{V}$ *be a finite set of variables with joint distribution* $P_{\mathbf{V}}$ *and assume (A1) - (A4) hold. Further, assume the general assumptions of the PC algorithm hold, i.e., causal faithfulness and causal Markov condition, see [46]. Let* $\hat{\mathcal{G}}_{CPDAG,n}(\alpha_n)$ *be the estimated CPDAG of the PC algorithm and* $\mathcal{G}_{CPDAG}$ *the CPDAG of the true underlying DAG* $\mathcal{G}$*. Then, for* $\alpha_n = \frac{1}{1+M_{perm,n}}$ *with* $M_{perm,n} \to \infty$ *as* $n \to \infty$*,*

$$\lim_{n\to\infty} \mathbb{P}_{P_{\mathbf{V}}} \left(\hat{\mathcal{G}}_{CPDAG,n}(\alpha_n) = \mathcal{G}_{CPDAG} \right) = 1. \tag{3}$$

The idea of the proof is to consider wrongly detected edges due to incorrect CI decisions and show that they can be controlled asymptotically. For detailed proof and more information on causal discovery, see Appendix C. As the upper bound on the errors is general for constraint-based methods, the consistency statement of Thm. 3 holds for modified versions of the PC algorithm, e.g., its order-independent version PC-stable [8], too. Hence, `mCMIkNN` for constraint-based causal discovery allows consistently estimating the $\mathcal{G}_{CPDAG}$ for $n \to \infty$.

5 Empirical Evaluation

We consider the mixed additive noise model (MANM) (Sect. 5.1) to synthetically examine `mCMIkNN`'s robustness (Sect. 5.2). Further, we compare `mCMIkNN`'s empirical performance against state-of-the-art competitors regarding CI decisions (Sect. 5.3), causal discovery (Sect. 5.4), and in a real-world scenario (Sect. 5.5).

5.1 Synthetic Data Generating

We generate synthetic data according to the MANM [24]. Hence, for all $X \in \mathbf{V}$, let X be generated from its J discrete parents $\mathcal{P}^{dis}(X) \subseteq \mathbf{V} \setminus X$, where $J := \#\mathcal{P}^{dis}(X)$, its K continuous parents $\mathcal{P}^{con}(X) \subseteq \mathbf{V} \setminus X$, where $K := \#\mathcal{P}^{con}(X)$, and (continuous or discrete) noise term N_X according to $X = \frac{1}{J}\sum_{j=1,\dots,J} f_j(Z_j) + (\sum_{k=1,\dots,K} f_k(Z_k)) \bmod d_X + N_X$ with appropriately defined functions f_j, f_k between $\mathbb{Z}$ and $\mathbb{R}$. Hence, by construction (A1) and (A2) hold true for all combinations of $X, Y, Z \subseteq \mathbf{V}$. For experimental evaluation, we generate CGMs that either directly induce CI characteristics between variables X and Y conditioned on $Z = \{Z_1, \dots, Z_{d_Z}\}$, d_Z between 1 and 7, (see Sect. 5.2 - 5.3) or are randomly generated with between 10 to 30 variables and varying densities between 0.1 and 0.4 (see Sect. 5.4). Moreover, we consider different ratios of discrete variables between 0 and 1. We consider the cyclic model with $d_X \in \{2, 3, 4\}$ for discrete X, and continuous functions that are equally drawn from $\{id(\cdot), (\cdot)^2, cos(\cdot)\}$. Note that we scale the parents' signals to reduce the

noise for subsequent variables avoiding high varsortability [41], and max-min normalize all continuous variables. For more information on the MANM and all parameters used for synthetic data generation, see Appendix D.1.

5.2 Calibration and Robustness of `mCMIkNN`

We provide recommendations for calibrating `mCMIkNN` and show its robustness, i.e., the ability to control type I and II errors in the finite case. Therefore, we restrict our attention to two simple CGMs $\mathcal{G}$ with variables $\mathbf{V} = (X, Y, Z_1, \ldots, Z_{d_Z})$, where first, X and Y have common parents $Z = \{Z_1, \ldots, Z_{d_Z}\}$ in $\mathcal{G}$, i.e., $H_0 : X \perp\!\!\!\perp Y \,|\, Z$, and second, there exists an additional edge connecting X and Y in $\mathcal{G}$, i.e., $H_1 : X \not\perp\!\!\!\perp Y | Z$. Accordingly, we generate the data using the MANM model with parameters described in Sect. 5.1.

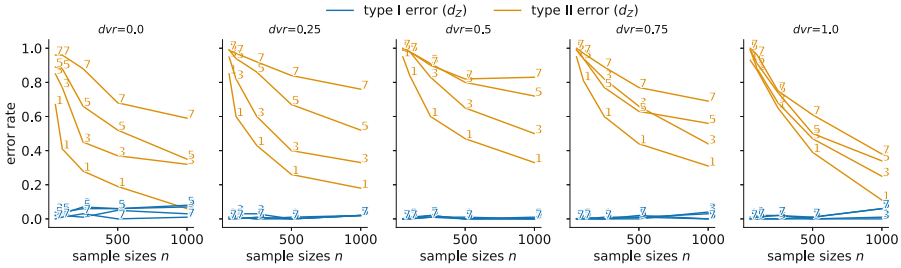

Fig. 1. Type I and II error rates of `mCMIkNN` for different dimensions $d_Z \in \{1,3,5,7\}$ of Z (smaller better) given varying sample sizes n for settings with different discrete variable ratios from $dvr = 0.0$, i.e., continuous (left), to $dvr = 1.0$, i.e., discrete (right).

Calibration: We evaluate the accuracy of CI decisions for different combinations of k_{CMI} and k_{perm} by comparing the area under the receiver operating curve (ROC AUC), as it provides a balanced measure of type I and type II errors. In particular, we examine different combinations of k_{CMI} and k_{perm} in settings with varying $d_Z \in \{1, 3, 5, 7\}$, discrete variable ratios $dvr \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and sample sizes n ranging from 50 to 1 000. Note, we set $\alpha = 0.05$ and $M_{perm} = 100$, cf. [14]. We find that small values of k_{CMI} and k_{perm} are sufficient to calibrate the CI test while not affecting accuracy much for the finite case, such that we set $k_{CMI} = 25$ and $k_{perm} = 5$ in the subsequent experiments. Note that Appendix D.2 provides detailed evaluation results. Moreover, for more information on all parameters, see Appendix A.

Robustness: We evaluate `mCMIkNN`'s robustness regarding validity and power in the finite case by examining the type I and II error rates as depicted in Fig. 1. In particular, we see that `mCMIkNN` is able to control type I errors for all discrete variable ratios dvr and sizes of the conditioning sets d_Z (cf. Appendix D.3). Moreover, the type II error rates decrease for an increasing number of samples

n. Hence, mCMIkNN achieves non-trivial power, particularly for small sizes of the conditioning sets d_Z. In this context, higher type II errors in the case of higher dimensions d_Z point out that mCMIkNN suffers from the curse of dimensionality, cf. Cor. 3. In summary, the empirical results are in line with the theoretical results on the asymptotic type I and II error control, cp. Thm. 1 and Thm. 2.

5.3 Conditional Independence Testing

Next, we compare mCMIkNN's empirical performance to state-of-the-art CI tests valid for mixed discrete-continuous data. We chose a likelihood ratio test assuming conditional Gaussianity (CG) [1], a discretization-based approach, where we discretize continuous variables before applying Pearson's χ^2 test (discχ²), a non-parametric CI test based upon adaptive histograms (aHistχ²) [36], and a non-parametric kernel-based CI test (KCIT) [52]. In this experiment, we again consider the two CGMs used for the calibration in Sect. 5.2 and examine the respective ROC AUC scores from 20 000 CI decisions ($\alpha = 0.01$) in Fig. 2.

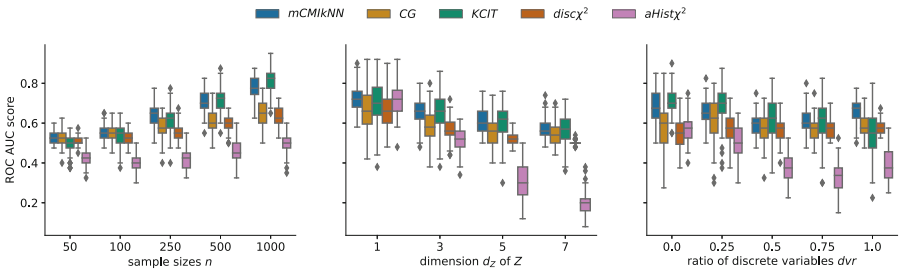

Fig. 2. ROC AUC scores (higher better) of 20 000 CI decisions of the CI tests mCMIkNN, CG, KCIT, discχ², and aHistχ² with varying sample sizes n (left), dimensions of the conditioning sets d_Z (center), and ratios of discrete variables dvr (right) (Note, we limited the execution time to 10 min per CI test (Approx. 4 900 runs of aHistχ² exceeded this time. Thus, aHistχ² is excluded for causal discovery).).

We compare the CI test's performance for various sample sizes (Fig. 2 left), sizes of conditioning sets d_Z (center), and ratios of discrete variables (right). While the ROC AUC scores of all CI tests increase as n grows (left), mCMIkNN outperforms all competitors, particularly for small sizes, e.g., $n \leq 500$. With increasing sample sizes, the performance of KCIT catches up to ROC AUC scores of mCMIkNN, cf. $n = 1\,000$. For an increasing size of the conditioning sets d_Z (center), we observe that all methods suffer from the curse of dimensionality, while mCMIkNN achieves higher ROC AUC scores than the competitors. Moreover, mCMIkNN achieves the highest ROC AUC independent of the ratio of discrete variables dvr (right), only beaten by KCIT for some dvr's. For a detailed evaluation and an examination of type I and II errors, see Appendix D.4.

5.4 Causal Discovery

We evaluate the consistency of causal discovery using the PC-stable algorithm from [8] ($\alpha = 0.05$ with $M_{perm} = 100$) to estimate $\mathcal{G}_{CPDAG}$ of the DAG $\mathcal{G}$ generated according to Sect. 5.1. We examine the F1 scores [7] of erroneously detected edges in the skeletons of $\hat{\mathcal{G}}_{CPDAG,n}(0.05)$ estimated with PC-stable using the respective CI tests in comparison to the true skeleton of $\mathcal{G}$, see Fig. 3. While F1 grows for all methods as n increases, `mCMIkNN` outperforms the competitors (left). Further, `mCMIkNN` achieves the highest F1 scores for high discrete variables ratios (center left). In this context, F1 scores are balanced towards type I errors, crucial in causal discovery. Further, constraint-based causal discovery requires higher sample sizes for consistency due to the multiple testing problem [17,46]. All methods suffer from the curse of dimensionality, i.e., a decreasing F1 score for increasing densities (center right) and numbers of variables (right) which yields larger conditioning sizes d_Z. For more information, see Appendix D.6.

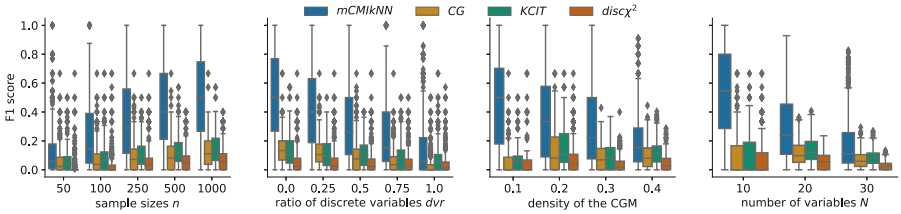

Fig. 3. F1 scores (higher better) of PC-stable with CI tests `mCMIkNN`, `CG`, `KCIT`, and `discχ²` computed over 3 000 CGMs for varying the sample sizes n, discrete variable ratios dvr, densities of CGMs, and numbers of variables N (left to right)[2].

5.5 Real-World Scenario: Discrete Manufacturing

Finally, we apply `mCMIkNN` in causal discovery on real-world manufacturing data. Therefore, we consider a simplified discrete manufacturing process whose underlying causal structures are confirmed by domain experts. In particular, we consider quality measurements Q_{con} and rejections R_{con} within a configuration phase used for adjustment of the processing speed S_{con} to reduce the number of rejected goods R_{prod} within a production phase. Besides these causal structures for configuration, rejections within the production phase R_{prod} vary given the corresponding locality within one of nine existing units U. In contrast to commonly applied discretization-based approaches, cf. [20], an experimental evaluation shows that `mCMIkNN` covers more of the CI characteristics present in the mixed discrete-continuous real-world data, hence, yields better estimates of causal structures when used in constraint-based causal discovery, F1 = 0.57 for `mCMIkNN` vs. F1 = 0.4 for `discχ²`. For additional details, see Appendix E.

6 Conclusion

We addressed the problem of testing CI in mixed discrete-continuous data and its application in causal discovery. We introduced the non-parametric CI test `mCMIkNN`, and showed its validity and power theoretically and empirically. We demonstrated that `mCMIkNN` outperforms state-of-the-art approaches in the accuracy of CI decisions, particularly for low sample sizes.

While mild assumptions simplify the application of `mCMIkNN` in practice, we cannot derive bounds on type I and II error control for the finite case as provided in [28], but the empirical results show that `mCMIkNN` is robust in the finite case, too. These bounds can be achieved by considering stronger assumptions, such as lower bounds on probabilities for discrete values, cf. [2,28], or smoothness assumptions for continuous variables, cf. [4,53]. Further, the current implementation of `mCMIkNN` is restricted to metric spaces. To extend the implementation to categorical variables, an isometric mapping into the metric space can be examined, cf. [37]. Note that kNN methods are not invariant regarding the scaling of variables, and their computational complexity yields long runtimes, particularly for large sample sizes. For an evaluation of runtimes, see Appendix D.5. We consider parallel execution strategies to speed up the computation, e.g., parallelizing the execution of M_{perm} permutations in Algorithm 2, cf. [43], or using GPUs [21].

References

1. Andrews, B., Ramsey, J., Cooper, G.F.: Scoring bayesian networks of mixed variables. Int. J. Data Sci. Analytics **6**(1), 3–18 (2018)
2. Antos, A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Struct. Algorithms **19**(3–4), 163–193 (2001)
3. Baba, K., Shibata, R., Sibuya, M.: Partial correlation and conditional correlation as measures of conditional independence. Aust. N. Z. J. Stat. **46**(4), 657–664 (2004)
4. Berrett, T.B., Wang, Y., Barber, R.F., Samworth, R.J.: The conditional permutation test for independence while controlling for confounders. J. Roy. Stat. Soc. B (Statistical Methodology) **82**(1), 175–197 (2020)
5. Bradley, J.V.: Distribution-Free Statistical Tests. Prentice-Hall, Inc. XII, Englewood Cliffs, N. J. (1968)
6. Cabeli, V., Verny, L., Sella, N., Uguzzoni, G., Verny, M., Isambert, H.: Learning clinical networks from medical records based on information estimates in mixed-type data. PLoS Comput. Biol. **16**(5), 1–19 (2020)
7. Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K.S., Liu, H.: Evaluation methods and measures for causal learning algorithms. IEEE Trans. Artif. Intell. **3**, 924–943 (2022)
8. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. **15**(116), 3921–3962 (2014)
9. Cui, R., Groot, P., Heskes, T.: Copula PC algorithm for causal discovery from mixed data. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) ECML PKDD 2016. LNCS (LNAI), vol. 9852, pp. 377–392. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46227-1_24

10. Cui, R., Groot, P., Schauer, M., Heskes, T.: Learning the causal structure of copula models with latent variables. In: Globerson, A., Silva, R. (eds.) Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI, pp. 188–197 (2018)
11. Dawid, A.P.: Conditional independence. Encycl. stat. sci. update **2**, 146–153 (1998)
12. Deckert, A.C., Kummerfeld, E.: Investigating the effect of binning on causal discovery. In: Proceedings of 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2574–2581 (2019)
13. Edwards, D.: Introduction to Graphical Modelling. Springer (2012)
14. Ernst, M.D.: Permutation methods: a basis for exact inference. Stat. Sci. **19**(4), 676–685 (2004)
15. Frenzel, S., Pompe, B.: Partial mutual information for coupling analysis of multivariate time series. Phys. Rev. Lett. **99**(20), 204101 (2007)
16. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 5988–5999 (2017)
17. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Front. genetics **10**, 524 (2019)
18. Gray, R.M.: Entropy and Information Theory. Springer (2011)
19. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: problems and methods. ACM Comput. Surv. **53**(4), 1–37 (2020)
20. Hagedorn, C., Huegle, J., Schlosser, R.: Understanding unforeseen production downtimes in manufacturing processes using log data-driven causal reasoning. J. Intell. Manuf. **33**(7), 2027–2043 (2022)
21. Hagedorn, C., Lange, C., Huegle, J., Schlosser, R.: GPU acceleration for information-theoretic constraint-based causal discovery. In: Proceedings of The KDD 2022 Workshop on Causal Discovery, pp. 30–60 (2022)
22. Higgins, J.J.: An Introduction to Modern Nonparametric Statistics. Brooks/Cole Pacific Grove, CA (2004)
23. Huang, T.M.: Testing conditional independence using maximal nonlinear conditional correlation. Ann. Stat. **38**(4), 2047–2091 (2010)
24. Huegle, J., Hagedorn, C., Boehme, L., Poerschke, M., Umland, J., Schlosser, R.: MANM-CS: data generation for benchmarking causal structure learning from mixed discrete-continuous and nonlinear data. In: WHY-21 @ NeurIPS 2021 (2021)
25. Huegle, J., Hagedorn, C., Uflacker, M.: How causal structural knowledge adds decision-support in monitoring of automotive body shop assembly lines. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 5246–5248 (2020)
26. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. **8**, 613–636 (2007)
27. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. J. Stat. Softw. **47**(11), 1–26 (2012)
28. Kim, I., Neykov, M., Balakrishnan, S., Wasserman, L.: Local permutation tests for conditional independence. Ann. Stat. **50**(6), 3388–3414 (2022)
29. Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. Probl. Inf. Transm. **23**(2), 9–16 (1987)
30. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**(6), 066138 (2004)
31. Lehmann, E.L., D'Abrera, H.J.M.: Nonparametrics: Statistical Methods Based on Ranks (1975)

32. Li, C., Fan, X.: On nonparametric conditional independence tests for continuous variables. Wiley Interdiscip. Rev. Comput. Stat. **12**(3) (2020)
33. Malinsky, D., Danks, D.: Causal discovery algorithms: a practical guide. Philos Compass **13**(1), e12470 (2018)
34. Mandros, P., Kaltenpoth, D., Boley, M., Vreeken, J.: Discovering functional dependencies from mixed-type data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1404–1414 (2020)
35. Margaritis, D.: Distribution-free learning of bayesian network structure in continuous domains. In: Proceedings of the National Conference on Artificial Intelligence, pp. 825–830. AAAI (2005)
36. Marx, A., Yang, L., van Leeuwen, M.: Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pp. 387–395 (2021)
37. Mesner, O.C., Shalizi, C.R.: Conditional mutual information estimation for mixed, discrete and continuous data. IEEE Trans. Inf. Theory **67**(1), 464–484 (2021)
38. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, 1st edn. (2000)
39. Pearson, K.: X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond. Edinb. Dublin Philos. Mag. J. Sci. **50**(302), 157–175 (1900)
40. Raghu, V.K., Poon, A., Benos, P.V.: Evaluation of causal structure learning methods on mixed data types. In: Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, vol. 92, pp. 48–65 (2018)
41. Reisach, A., Seiler, C., Weichwald, S.: Beware of the simulated dag! causal discovery benchmarks may be easy to game. In: Advances in Neural Information Processing Systems, vol. 34, pp. 27772–27784 (2021)
42. Rohekar, R.Y., Nisimov, S., Gurwicz, Y., Novik, G.: Iterative causal discovery in the possible presence of latent confounders and selection bias. Adv. Neural. Inf. Process. Syst. **34**, 2454–2465 (2021)
43. Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: International Conference on Artificial Intelligence and Statistics, pp. 938–947. PMLR (2018)
44. Scutari, M.: Learning bayesian networks with the bnlearn R package. J. Stat. Softw. **35**, 1–22 (2010)
45. Shah, R.D., Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. Ann. Stat. **48**(3), 1514–1538 (2020)
46. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search. MIT Press, Adaptive Computation and Machine Learning (2000)
47. Strobl, E.V.: A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. Int. J. Data Sci. Analytics **8**(1), 33–56 (2019)
48. Tsagris, M., Borboudakis, G., Lagani, V., Tsamardinos, I.: Constraint-based causal discovery with mixed data. Int. J. Data Sci. Analytics **6**(1), 19–30 (2018)
49. Tsamardinos, I., Borboudakis, G.: Permutation testing improves bayesian network learning. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) Machine Learning and Knowledge Discovery in Databases, pp. 322–337. Springer, Berlin Heidelberg, Berlin, Heidelberg (2010)
50. Yu, K., et al.: Causality-based feature selection: methods and evaluations. ACM Comput. Surv. **53**(5), 1–36 (2020)

51. Zan, L., Meynaoui, A., Assaad, C.K., Devijver, E., Gaussier, E.: A conditional mutual information estimator for mixed data and an associated conditional independence test. Entropy **24**(9), 1234 (2022)
52. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pp. 804–813 (2011)
53. Zhao, P., Lai, L.: Analysis of KNN information estimators for smooth distributions. IEEE Trans. Inf. Theory **66**(6), 3798–3826 (2019)